Python Data Products

Course 1: Basics

Lecture: CSV and JSON files

Learning objectives

In this lecture we will...

- Demonstrate the CSV/TSV and JSON formats
- Compare the main advantages and disadvantages of both formats

CSV and JSON

CSV and JSON are two formats that are easy to read and manipulate in Python

- We'll work through two examples for much of this course:
  - https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt (TSV)
  - https://www.yelp.com/dataset/download (JSON)

CSV and JSON

First let's look at the Amazon dataset:

- https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt
- We'll look at data from the "Gift Card" category

Concept: CSV

- CSV is a simple format that allows us to store **tabular** data
- It is a human-readable format, meaning it can easily be read or written via a text-editor or spreadsheet application

E.g. CSV (or rather TSV) from Amazon

```
marketplace    customer_id    review_id    product_id    product_parent    product_title    product_category    star_rating    help
US    24371595    R27ZP1F1CD0C3Y    B004LLIL5A    346014806    Amazon eGift Card - Celebrate    Gift Card    5    0    0    N    Y
US    42489718    RJ7RSBCHUDNNE    B004LLIKVU    473048287    Amazon.com eGift Cards    Gift Card    5    0    0    N    Y    Gift
US    861463    R1HVYBSKLQJI5S    B00IX1I3G6    926539283    Amazon.com Gift Card Balance Reload    Gift Card    5    0    0    N
US    25283295    R2HAXF0IIYQBIR    B00IX1I3G6    926539283    Amazon.com Gift Card Balance Reload    Gift Card    1    0    0    N
US    397970    RNYLPX611NB7Q    B005ESMGV4    379368939    Amazon.com Gift Cards, Pack of 3 (Various Designs)    Gift Card    5
US    18513645    R3ALA9XXMBEDZR    B004KNWWU4    326384774    Amazon Gift Card - Print - Happy Birthday (Birds)    Gift Card    5
US    22484620    R3R8PHAVJFTPDF    B004LLIKVU    473048287    Amazon.com eGift Cards    Gift Card    5    0    0    N    Y    Five
US    14765851    R18WWEK8OIXE30    BT00CTP2EE    775486538    Amazon.com Gift Card in a Greeting Card (Various Designs)    Gift Ca
US    18751931    R1EGUNQON2J277    B004LLIKVU    473048287    Amazon.com eGift Cards    Gift Card    1    0    0    N    Y    One
US    15100528    R21Z4M4L98CPU2    B004W8D102    595099956    Amazon Gift Card - Print - Amazon Boxes    Gift Card    5    0    0
US    3559726    R6JH7A117FHFA    B004LLIKVU    473048287    Amazon.com eGift Cards    Gift Card    5    0    0    N    Y    Five S
US    23413911    R1XZHS8M1GCGI7    B004KNWWU4    326384774    Amazon Gift Card - Print - Happy Birthday (Birds)    Gift Card    5
US    2026222    R1DAI0N03SKRJN    B004LLIKVU    473048287    Amazon.com eGift Cards    Gift Card    5    1    1    N    Y    Five
US    32956435    R2F6SKZOEYQRU3    B004LLIKVU    473048287    Amazon.com eGift Cards    Gift Card    5    0    0    N    N    Five
US    20241560    RIBOP6OEAZA47    B00H5BNLUS    637715957    Amazon eGift Card - Hoops and Yoyo Thank You Very Much (Animated) [Ha
US    10670435    R15H8E7WD6XD29    B004KNWX6C    763371347    Amazon Gift Card - Print - Celebrate    Gift Card    5    0    0
US    48872127    RVN4P3RU4F8IE    BT00CTOYC0    506740729    Amazon.com $15 Gift Card in a Greeting Card (Amazon Surprise Box Desi
US    460630    RCS8F9JCAAXC7    B004LLIKVU    473048287    Amazon.com eGift Cards    Gift Card    4    0    0    N    Y    Four St
US    41238378    R6811C4E7UYL2    B00H5BMH44    81025991    Amazon eGift Card - Hoops and Yoyo Cake Face (Animated) [Hallmark]
```

TSV example

From Amazon's public dataset ("Gift Card" category – amazon_reviews_us_Gift_Card_v1_00.tsv.gz):
https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating	
US	24371595	R27ZP1F1CD0C3Y	B004LLIL5A	346014806	Amazon eGift Card - Celebrate	Gift Card	5	0	0	N
US	42489718	RJ7RSBCHUDNE	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	0	0	N	Y	G
US	861463	R1HVYBSKLQJI5S	B00IX1I3G6	926539283	Amazon.com Gift Card Balance Reload	Gift Card	5	0	0
US	25283295	R2HAXF0IIYQ3IR	B00IX1I3G6	926539283	Amazon.com Gift Card Balance Reload	Gift Card	1	0	0
US	397970	RNYLPX611NB7Q	B005ESMGV4	379368939	Amazon.com Gift Cards, Pack of 3 (Various Designs)	Gift Card
US	18513645	R3ALA9XXMBEDZR	B004KNWWU4	326384774	Amazon Gift Card - Print - Happy Birthday (Birds)	Gift Card
US	22484620	R3R8PHAVJFTPDF	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	0	0	N	Y
US	14765851	R18WWEK8OIXE30	BT00CTP2EE	775486538	Amazon.com Gift Card in a Greeting Card (Various Designs)	Gif
US	18751931	R1EGUNQON2J277	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	1	0	0	N	Y
US	15100528	R21Z4M4L98CPU2	B004W8D102	595099956	Amazon Gift Card - Print - Amazon Boxes	Gift Card	5	0
US	3559726	R6JH7A117FHFA	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	0	0	N	Y	Fi
US	23413911	R1XZHS8M1GCGI7	B004KNWWU4	326384774	Amazon Gift Card - Print - Happy Birthday (Birds)	Gift Card
US	2026222	R1DAI0N03SKRJN	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	1	1	N	Y	F
US	32956435	R2F6SK[illegible]	[illegible]KVU	473048287	Amazon.com eGift Cards	Gift Card	5	0	0	N	N
US	20241560	RIBOP[illegible]	[illegible]US	637715957	Amazon eGift Card - Hoops and Yoyo Thank You Very Much (Animated)
US	10670435	R15H8[illegible]	[illegible]X6C	763371347	Amazon Gift Card - Print - Celebrate	Gift Card	5	0	0
US	48872127	RVN4P[illegible]	[illegible]C0	506740729	Amazon.com $15 Gift Card in a Greeting Card (Amazon Surprise Box
US	460630	RCS8F9J[illegible]	[illegible]	473048287	Amazon.com eGift Cards	Gift Card	4	0	0	N	Y	Fou
US	41238378	R6811C4E7UYL2	B00H5BMH44	81025991	Amazon eGift Card - Hoops and Yoyo Cake Face (Animated) [Hallmark]

Data are separated by ***tabs*** (tsv) or ***commas*** (csv)

TSV example

From Amazon's public dataset ("Gift Card" category – amazon_reviews_us_Gift_Card_v1_00.tsv.gz):
https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating	
US	24371595	R27ZP1F1CD0C3Y	B004LLIL5A	346014806	Amazon eGift Card - Celebrate	Gift Card	5	0	0	N
US	42489718	RJ7RSBCHUDNNE	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	0	0	N	Y	G
US	861463	R1HVYBSKLQJI5S	B00IX1I3G6	926539283	Amazon.com Gift Card Balance Reload	Gift Card	5	0	0	
US	25283295	R2HAXF0IIYQBIR	B00IX1I3G6	926539283	Amazon.com Gift Card Balance Reload	Gift Card	1	0	0	
US	397970	RNYLPX611NB7Q	B005ESMGV4	379368939	Amazon.com Gift Cards, Pack of 3 (Various Designs)	Gift Card				
US	18513645	R3ALA9XXMBEDZR	B004KNWWU4	326384774	Amazon Gift Card - Print - Happy Birthday (Birds)	Gift Card				
US	22484620	R3R8PHAVJFTPDF			mazon.com eGift Cards	Gift Card	5	0	0	N	Y	
US	14765851	R18WWEK8OIXE30			mazon.com Gift Card in a Greeting Card (Various Designs)	Gif				
US	18751931	R1EGUNQON2J277			mazon.com eGift Cards	Gift Card	1	0	0	N	Y	
US	15100528	R21Z4M4L98CPU2			mazon Gift Card - Print - Amazon Boxes	Gift Card	5	0		
US	3559726	R6JH7A117FHFA			zon.com eGift Cards	Gift Card	5	0	0	N	Y	Fi
US	23413911	R1XZHS8M1GCGI7	B004KNWWU4	326384774	Amazon Gift Card - Print - Happy Birthday (Birds)	Gift Card				
US	2026222	R1DAI0N03SKRJN	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	1	1	N	Y	F
US	32956435	R2F6SKZOEYQRU3	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	0	0	N	N	
US	20241560	RIBOP6OEAZA47	B00H5BNLUS	637715957	Amazon eGift Card - Hoops and Yoyo Thank You Very Much (Animated)					
US	10670435	R15H8E7WD6XD29	B004KNWX6C	763371347	Amazon Gift Card - Print - Celebrate	Gift Card	5	0	0	
US	48872127	RVN4P3RU4F8IE	BT00CTOYC0	506740729	Amazon.com $15 Gift Card in a Greeting Card (Amazon Surprise Box					
US	460630	RCS8F9JCAAXC7	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	4	0	0	N	Y	Fou
US	41238378	R6811C4E7UYL2	B00H5BMH44	81025991	Amazon eGift Card - Hoops and Yoyo Cake Face (Animated) [Hallmark]					

The first row is the ***header***, which indicates what value each field represents

TSV example

From Amazon's public dataset ("Gift Card" category – amazon_reviews_us_Gift_Card_v1_00.tsv.gz):
https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating	
US	24371595	R27ZP1F1CD0C3Y	B004LLIL5A	346014806	Amazon eGift Card - Celebrate	Gift Card	5	0	0	N
US	42489718	RJ7RSBCHUDNNE	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	0	0	N	Y	G
US	861463	R1HVYBSKLQJI5S	B00IX1I3G6	926539283	Amazon.com Gift Card Balance Reload	Gift Card	5	0	0	
US	25283295	R2HAXF0IIYQBIR	B00IX1I3G6	926539283	Amazon.com Gift Card Balance Reload	Gift Card	1	0	0	
US	397970	RNYLPX611NB7Q	B005ESMGV4	379368939	Amazon.com Gift Cards, Pack of 3 (Various Designs)	Gift Card				
US	18513645	R3ALA9XXMBEDZR	B004KNWWU4	326384774	Amazon Gift Card - Print - Happy Birthday (Birds)	Gift Card				
US	22484620	R3R8PHAVJFTPDF	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	0	0	N	Y	
US	14765851	R18WWEK8OIXE30	BT00CTP2EE	775486538	Amazon.com Gift Card in a Greeting Card (Various Designs)	Gif				
US	18751931	R1EGUNQON2J277	B004LLIKV		m eGift Cards	Gift Card	1	0	0	N	Y	
US	15100528	R21Z4M4L98CPU2	B004W8D1(		ft Card - Print - Amazon Boxes	Gift Card	5	0		
US	3559726	R6JH7A117FHFA	B004LLIKVU		eGift Cards	Gift Card	5	0	0	N	Y	Fi
US	23413911	R1XZHS8M1GCGI7	B004KNWWU		ft Card - Print - Happy Birthday (Birds)	Gift Card				
US	2026222	R1DAI0N03SKRJN	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	1	1	N	Y	F
US	32956435	R2F6SKZOEYQRU3	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	5	0	0	N	N	
US	20241560	RIBOP6OEAZA47	B00H5BNLUS	637715957	Amazon eGift Card - Hoops and Yoyo Thank You Very Much (Animated)					
US	10670435	R15H8E7WD6XD29	B004KNWX6C	763371347	Amazon Gift Card - Print - Celebrate	Gift Card	5	0	0	
US	48872127	RVN4P3RU4F8IE	BT00CTOYC0	506740729	Amazon.com $15 Gift Card in a Greeting Card (Amazon Surprise Box					
US	460630	RCS8F9JCAAXC7	B004LLIKVU	473048287	Amazon.com eGift Cards	Gift Card	4	0	0	N	Y	Fou
US	41238378	R6811C4E7UYL2	B00H5BMH44	81025991	Amazon eGift Card - Hoops and Yoyo Cake Face (Animated) [Hallmark]					

Each other row corresponds to a single product review from *Amazon*

TSV example

Can be a little easier to visualize if we align the columns vertically:

Note that the data is essentially ***tabular*** and is much like an *Excel* (or similar) spreadsheet

marketplace	customer_id	review_id	product_id	product_parent	product_title
US	24371595	R27ZP1F1CD0C3Y	B004LLIL5A	346014806	Amazon eGift Card - Celebrate
US	42489718	RJ7RSBCHUDNNE	B004LLIKVU	473048287	Amazon.com eGift Cards
US	861463	R1HVYBSKLQJI5S	B00IX1I3G6	926539283	Amazon.com Gift Card Balance Reload
US	25283295	R2HAXF0IIYQBIR	B00IX1I3G6	926539283	Amazon.com Gift Card Balance Reload
US	397970	RNYLPX611NB7Q	B005ESMGV4	379368939	Amazon.com Gift Cards, Pack of 3 (Various Designs)
US	18513645	R3ALA9XXMBEDZR	B004KNWWU4	326384774	Amazon Gift Card - Print - Happy Birthday (Birds)
US	22484620	R3R8PHAVJFTPDF	B004LLIKVU	473048287	Amazon.com eGift Cards
US	14765851	R18WWEK8OIXE30	BT00CTP2EE	775486538	Amazon.com Gift Card in a Greeting Card (Various D
US	18751931	R1EGUNQON2J277	B004LLIKVU	473048287	Amazon.com eGift Cards
US	15100528	R21Z4M4L98CPU2	B004W8D102	595099956	Amazon Gift Card - Print - Amazon Boxes
US	3559726	R6JH7A117FHFA	B004LLIKVU	473048287	Amazon.com eGift Cards
US	23413911	R1XZHS8M1GCGI7	B004KNWWU4	326384774	Amazon Gift Card - Print - Happy Birthday (Birds)
US	2026222	R1DAI0N03SKRJN	B004LLIKVU	473048287	Amazon.com eGift Cards
US	32956435	R2F6SKZOEYQRU3	B004LLIKVU	473048287	Amazon.com eGift Cards
US	20241560	RIBOP6OEAZA47	B00H5BNLUS	637715957	Amazon eGift Card - Hoops and Yoyo Thank You Very
US	10670435	R15H8E7WD6XD29	B004KNWX6C	763371347	Amazon Gift Card - Print - Celebrate
US	48872127	RVN4P3RU4F8IE	BT00CTOYC0	506740729	Amazon.com $15 Gift Card in a Greeting Card (Amazo
US	460630	RCS8F9JCAAXC7	B004LLIKVU	473048287	Amazon.com eGift Cards
US	41238378	R6811C4E7UYL2	B00H5BMH44	81025991	Amazon eGift Card - Hoops and Yoyo Cake Face (Anim

TSV example

Let's look at some more columns of the data:

product_category	star_rating	helpful_votes	total_votes	vine	verified_purchase	review_headline	review_body	rev
Gift Card	5	0	0	N	Y	Five Stars	Great birthday gi	
Gift Card	5	0	0	N	Y	Gift card for the greatest	selection of iter	
Gift Card	5	0	0	N	Y	Five Stars	Good 2015-08-3	
Gift Card	1	0	0	N	Y	One Star	Fair 2015-08-3	
Gift Card							I can't believe ho	
Gift Card							casion! Perfect	
Gift Card	5	0	0	N	Y	Five Stars	excelent 2015-	
Gift Card	5	0	0	N	Y	Five Stars	Great and Safe Gi	
Gift Card	1	0	0	N	Y	One Star	What?????????	
Gift Card	5	0	0	N	Y	Five Stars	This was just too	
Gift Card	5	0	0	N	Y	Five Stars	Bien 2015-08-3	
Gift Card	5	1	1	N	Y	Always good	Easy to print fro	
Gift Card	5	1	1	N	Y	Five Stars	Amazing with 10 d	
Gift Card	5	0	0	N	N	Five Stars	Remember Matthew	
Gift Card	5	1	1	N	Y	Five Stars	good 2015-08-3	
Gift Card	5	0	0	N	Y	Five Stars	Awesome way to se	
Gift Card	5	0	0	N	Y	Quick Solution for Forgotten Occasion	I	
Gift Card	4	0	0	N	Y	Four Stars	Good gift. Easy t	
Gift Card	5	0	0	N	Y	Satisfied customer	Satisfied as usua	

What was the rating, how many helpful votes were received (out of how many), was the purchase verified, etc.

Concept: JSON

- CSV/TSV format is **simple and effective**, but **limited** in the types of data that can be stored
- I.e., it is limited to **tabular** data
- For example, how would you store
  - A business's opening hours (e.g. in *Yelp's* data)?
  - A set of categories for a product?
  - Playlist entries in the "million playlist dataset"? (https://labs.spotify.com/2018/05/30/introducing-the-million-playlist-dataset-and-recsys-challenge-2018/)
- Fields best represented as **lists** or **sets** are inconvenient to store as CSV/TSV entries
- **JSON** attempts to address this by allowing for more general structured data to be represented

Concept: JSON

- See e.g. yelp's dataset: https://www.yelp.com/dataset/download
- Let's look at the first line of the "business.json" file:


```
{'business_id': 'FYWN1wneV18bWNgQjJ2GNg', 'attributes':
{'BusinessAcceptsCreditCards': True, 'AcceptsInsurance': True,
'ByAppointmentOnly': True}, 'longitude': -111.9785992,
'state': 'AZ', 'address': '4855 E Warner Rd, Ste B9',
'neighborhood': '', 'city': 'Ahwatukee', 'hours': {'Tuesday':
'7:30-17:00', 'Wednesday': '7:30-17:00', 'Thursday': '7:30-
17:00', 'Friday': '7:30-17:00', 'Monday': '7:30-17:00'},
'postal_code': '85044', 'review_count': 22, 'stars': 4.0,
'categories': ['Dentists', 'General Dentistry', 'Health &
Medical', 'Oral Surgeons', 'Cosmetic Dentists',
'Orthodontists'], 'is_open': 1, 'name': 'Dental by Design',
'latitude': 33.3306902}
```

JSON

Might look a little cleaner if we format it more carefully:


```
{
  'business_id': 'FYWN1wneV18bWNgQjJ2GNg',
  'attributes':
  {
    'BusinessAcceptsCreditCards': True,
    'AcceptsInsurance': True,
    'ByAppointmentOnly': True
  },
  'longitude': -111.9785992,
  'latitude': 33.3306902,
  'state': 'AZ',
  'address': '4855 E Warner Rd, Ste B9',
  'neighborhood': '',
  'city': 'Ahwatukee',
  'hours':
  {
    'Tuesday': '7:30-17:00',
    'Wednesday': '7:30-17:00',
    'Thursday': '7:30-17:00',
    'Friday': '7:30-17:00',
    'Monday': '7:30-17:00'
  },
  'postal_code': '85044',
  'review_count': 22,
  'stars': 4.0,
  'categories':
  ['Dentists', 'General Dentistry', 'Health & Medical', 'Oral Surgeons', 'Cosmetic Dentists', 'Orthodontists'],
  'is_open': 1,
  'name': 'Dental by Design'
}
```

JSON

Might look a little cleaner if we format it more carefully:

```
{
  'business_id': 'FYWN1wneV18bWNgQjJ2GNg',
  'attributes':
  {
    'BusinessAcceptsCreditCards': True,
    'AcceptsInsurance': True,
    'ByAppointmentOnly': True
  },
  'longitude': -111.9785992,
  'latitude': 33.3306902,
  'state': 'AZ',
  'address': '4855 E Warner Rd, Ste B9',
  'neighborhood': '',
  'city': 'Ahwatukee',
  'hours':
  {
    'Tuesday': '7:30-17:00',
    'Wednesday': '7:30-17:00',
    'Thursday': '7:30-17:00',
    'Friday': '7:30-17:00',
    'Monday': '7:30-17:00'
  },
  'postal_code': '85044',
  'review_count': 22,
  'stars': 4.0,
  'categories':
  ['Dentists', 'General Dentistry', 'Health & Medical', 'Oral Surgeons', 'Cosmetic Dentists', 'Orthodontists'],
  'is_open': 1,
  'name': 'Dental by Design'
}
```

Note that (unlike TSV), whitespace, newlines, or the **ordering of entries** don't change the **meaning** of the JSON string

JSON

Might look a little cleaner if we format it more carefully:


```
{
  'business_id': 'FYWN1wneV18bWNgQjJ2GNg',
  'attributes':
  {
    'BusinessAcceptsCreditCards': True,
    'AcceptsInsurance': True,
    'ByAppointmentOnly': True
  },
  'longitude': -111.9785992,
  'latitude': 33.3306902,
  'state': 'AZ',
  'address': '4855 E Warner Rd, Ste B9',
  'neighborhood': '',
  'city': 'Ahwatukee',
  'hours':
  {
    'Tuesday': '7:30-17:00',
    'Wednesday': '7:30-17:00',
    'Thursday': '7:30-17:00',
    'Friday': '7:30-17:00',
    'Monday': '7:30-17:00'
  },
  'postal_code': '85044',
  'review_count': 22,
  'stars': 4.0,
  'categories':
  ['Dentists', 'General Dentistry', 'Health & Medical', 'Oral Surgeons', 'Cosmetic Dentists', 'Orthodontists'],
  'is_open': 1,
  'name': 'Dental by Design'
}
```


Each value is either a
- String
- Boolean
- Number
- A list
- **Another JSON object!**

CSV/TSV vs. JSON

CSV/TSV:

Advantages:

+ Simple, human-readable format
+ Can be easily manipulated in "tabular" form – e.g can be read & modified using *Excel* or similar tools

Disadvantages:

- Cannot represent complex, flexible (I.e., non-tabular) data

JSON:

Advantages:

+ Allows for manipulation of complex, semi-structured data

Disadvantages:

- More difficult to explore and manipulate using "GUI" tools

Summary of concepts

- You should understand the **format** of json and csv files
- Understand the relative **merits** of both formats

On your own…

- Download the Amazon and Yelp datasets
- Try opening the (smaller) files in a text editor